A Model of Perceived Dynamic Range for HDR Images

Vedad Hulusic[a,*], Kurt Debattista[c], Giuseppe Valenzise[b], Frédéric Dufaux[b]

[a]*LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 46 Rue Barrault, 75013, Paris, France*
[b]*Laboratoire des Signaux et Systèmes (L2S, UMR 8506), CNRS - CentraleSupélec - Université Paris-Sud, 91192 Gif-sur-Yvette, France*
[c]*WMG, University of Warwick, CV4 7AL, Coventry, UK*

Abstract

For High Dynamic Range (HDR) content, the dynamic range of an image is an important characteristic in algorithm design and validation, analysis of aesthetic attributes and content selection. Traditionally, it has been computed as the ratio between the maximum and minimum pixel luminance, a purely objective measure; however, the human visual system's perception of dynamic range is more complex and has been largely neglected in the literature. In this paper, a new methodology for measuring perceived dynamic range (PDR) of chromatic and achromatic HDR images is proposed. PDR can benefit HDR in a number of ways: for evaluating inverse tone mapping operators and HDR compression methods; aesthetically; or as a parameter for content selection in perceptual studies. A subjective study was conducted on a data set of 36 chromatic and achromatic HDR images. Results showed a strong agreement across participants' allocated scores. In addition, a high correlation between ratings of the chromatic and achromatic stimuli was found. Based on the results from a pilot study, five objective measures (pixel-based dynamic range, image key, area of bright regions, contrast and colorfulness) were selected as candidates for a PDR predictor model; two of which have been found to be significant contributors to the model. Our analyses show that this model performs better than individual metrics for both achromatic and chromatic stimuli.

Keywords: High Dynamic Range, Perceived Dynamic Range, Subjective Evaluation, Predictor Model

1. Introduction

High dynamic range (HDR) technology [1, 2] enables the capture, storage, transmission and display of the full range of real-world lighting and colors, with a significant increase in precision when compared to traditional low dynamic range (LDR) imaging. One of HDR's main features is its ability to reproduce very bright and very dark portions of a scene concurrently. The span between these extrema in the brightness scale is commonly referred to as the *dynamic range* of a picture.

The dynamic range of image or video content is frequently reported by many HDR applications. It is typically computed as the ratio between the maximum and minimum pixel luminance of an image, which will be referred to as *pixel-based* dynamic range (DR) in this paper. Such a computation can be biased due to image noise or singularities, such as isolated pixels with extreme luminance values. Furthermore, such measures do not capture the complex behavior of the human visual system's (HVS) response and perception of lightness, including its intrinsic content-dependency [3]. The perceived dynamic range (PDR) of HDR content and its assessment in HDR conditions still remain unexplored. The accurate prediction of PDR would be important for a number of applications. It could be used to optimize and evaluate inverse tone mapping operators (ITMOs) [4, 5, 6], HDR compression methods and HDR reproduction systems [7]; it could be used for developing objective image quality measures and quantifying aesthetic attributes [8]; it provides an objective means to select content for HDR presentations and subjective studies [9]; and in general it would help to better understand lightness and color perception, by extending studies on the anchoring problem [3] to complex stimuli and HDR conditions.

To the authors' knowledge, this work is the first attempt to assess and predict the perceived dynamic range of HDR images under HDR conditions. Currently neither a standardized methodology nor an HDR data set with annotated measurements of this perceptual attribute exist. For this purpose, a subjective study with 23 participants was designed and conducted, using a set of 36 HDR images (chosen from a larger pool) with different characteristics and content semantics, including indoor/outdoor scenes, natural/man-made scenes and other variations. This work is an extended version of the pilot study that investigated only the achromatic stimuli [10]; this extension has increased the number of participants, added chromatic stimuli to the study and used the data to propose and evaluate a predictive model. While dynamic range is generally measured based only on the brightness of a picture, well-known color appearance phenomena such as Hunt or Helmholtz-Kohlrausch effects [11] tend to question this assumption and rather lean towards the hypothesis that dynamic range perception changes from achromatic to chromatic stimuli. Therefore, in this work, both the achromatic and chromatic images were used and the correlation of the subjective scores is investigated.

This paper makes the following contributions:

*Corresponding author
Email address: vedad.hulusic@telecom-paristech.fr (Vedad Hulusic)

- a subjectively annotated data set with PDR values, using complex, chromatic and achromatic stimuli and HDR viewing conditions (using an HDR display) was created [1];
- a novel test methodology for measuring perceived dynamic range, partially inspired by the subjective assessment methodology for video quality (SAMVIQ) [12] is proposed;
- based on the results of the study, the Pearson's correlations between mean opinion scores (MOS) and five image features, i.e., pixel-based dynamic range (correlation coefficient r = .87 achromatic, r = .84 chromatic), image key (r = -.60, r = -.61), area (modified to account for non-linearity; r = .87, r = .87), contrast (r = -.19, r = -.22) and colorfulness (r = -.47 chromatic only) were analyzed;
- the effect of chromatic information on perceived dynamic range was investigated and the relation between chromatic and achromatic quantified (r = .99);
- a model for predicting the perceived dynamic range for both achromatic (adjusted R^2 = .89) and chromatic (adjusted R^2 = .87) images has been proposed.

In the rest of the paper, the abbreviations and terms presented in Table 1 will be used.

Table 1: A list of abbreviations and terms used in the paper.

Term	Meaning
PDR	Perceived dynamic range
MDR	Dynamic range as predicted by the proposed regression model
HVS	Human visual system
iTMO	Inverse tone mapping operator
MOS	Mean opinion score
CSF	Contrast sensitivity function
IK	Image key
SI	Spatial information
C	Contrast
Col	Colorfulness
F(x,y)	F-ratio (test statistic used in ANOVA)
p	The probability value
r	Pearson's correlation coefficient
r_s	Spearman's rank correlation coefficient
W	Kendall's coefficient of concordance
R	The multiple correlation coefficient
R^2	The coefficient of determination
adj. R^2	The adjusted R^2 coefficient
B	The regression coefficient (unstandardized)
Beta	Standardized regression coefficient
SR	Standardized residuals
CD	Cook's distance
VIF	Variance inflation factor

2. Related Work

The study of perceived dynamic range shares some common features with perceived lightness, image contrast and image quality. Most of these psycho-perceptual theories lack sufficient validation with complex stimuli, and have never been tested in HDR conditions.

[1] The data set is available on the project website: http://pdr.lefca.net

2.1. Perceived Lightness

Lightness is a measure of relative brightness of elements of a scene, and has been extensively studied within the field of perceptual psychology. One of the most popular models in lightness perception is the intrinsic image model [13, 14, 15, 16]. This is typically based on a multi-component approach, which consists in factorizing the perceived scene as the interaction of different elements, e.g., surface reflectance, illumination and three-dimensional form or depth values. However, Gilchrist argued that these models are incomplete and proposed a new anchoring model [3], based on a combination of global and local anchoring of lightness values. On the whole, the anchoring models promote the idea that the perception of lightness is determined by the brightest patch of the scene. The human visual system then scales the rest to this maximum, generating an internal, scene-dependent scale of light and dark. Furthermore, Li and Gilchrist [17] observed that anchoring is affected by the relative *area* of the brightest patch.

The Retinex theory [18] attempts to generate an output that is most similar to what a human observer would perceive by looking at the real scene where an image was taken. Compared with the anchoring theories, the Retinex theory indirectly arrives to the same conclusion through a probabilistic formulation. This is achieved by averaging luminance values along paths of pixels originating from each point of the picture [19], while taking into account the relative distance between patches of different brightness.

2.2. Image Contrast

While image contrast and dynamic range have similar perceptual connotation, they describe different aesthetic attributes. Global contrast measures [20, 21, 22] model image sharpness, which is a fine-scale image feature, as the perceptual experiments by Haun and Peli [23] clearly display. On the contrary, the perceived dynamic range is a measure of the magnitude of the global difference in the perceived image brightness. This is closely related to the "tone" aesthetic attribute as defined in work by Aydin et al. [8]. However, in their work this attribute is computed as a variation of the pixel-based dynamic range and does not take into account other perceptual factors. In a series of studies by Calabria and Fairchild it has been shown that perceived image contrast for LDR images is a function of image lightness, chroma, and sharpness [24] and the model of perceived image contrast and observed preference data were proposed [25].

In their recent work, Vangorp et al. defined a model of local adaptation and use it to measure dynamic range as the ratio between the brightest and darkest image region in which people can still see some details [26]. The model predicts the maximum visible dynamic range for any given scene, based on both glare and local adaptation. The results revealed that the greatest decrease in perceived dynamic range, compared to the physical DR, occurs in darker scene regions due to glare. On the other hand, local adaptation causes a significant loss of dynamic range visibility in brighter parts of the scenes.

2.3. Image Quality

Image quality metrics are usually based on either predicting the visibility of distortions [27, 28] or magnitude of aesthetic attributes [8]. While such metrics usually do not consider perceived dynamic range explicitly, there is evidence that the latter is highly related with the overall image quality. In the study by Akyüz et al., when two versions of the same image were compared, observers generally preferred the higher dynamic range one [29].

Therefore, assessing the perceived dynamic range of HDR pictures and videos can make a great contribution in design and evaluation of HDR methods, for example, inverse tone mapping operators (iTMOs) [4, 30], where the dynamic range of LDR content is expanded for displaying on an HDR display. Most of the current approaches to enhance the perception of dynamic range are based on heuristics. Meylan et al. [31] showed that, when expanding the dynamic range of LDR content, specular highlights have to be allocated a significant range. Similarly, other studies tend to boost the brightest pixels of the LDR scene in order to "approximate the visceral response associated with the higher contrast and overall brightness in the original scene" [5, 4]. The results of the subjective experiment conducted in this paper provide a groundtruth for designing more perceptually meaningful dynamic range metrics.

Furthermore, as pointed out by Narwaria et al. [9], having a perceptually annotated HDR image dataset with MOS values is necessary for both HDR source content selection in subjective studies and for testing HDR processing algorithms. However, in that work the proposed objective measure is tailored to a dynamic range reduction task, such as in the case of tone mapping evaluation. Moreover, no formal subjective evaluation is proposed to verify the perceptual relevance of that method. Predictive models, if successful, can serve the same role as annotated databases without requiring time-consuming data collection.

3. Motivation and Methodology

Computational metrics that predict PDR could allow for automatic organization, ordering and presentation of HDR content. This work's goal is to analyze how humans perceive dynamic range of HDR content on complex scenes, and be able to predict the PDR of a scene automatically. The research method employed involves data collection and analysis of HDR content from real-world HDR images. This data is compared with objective measurements; these are subsequently used to build a PDR model. This section describes the choice of objective metrics and describes the experimental methodology.

3.1. Objective Evaluation

The objective measurements proposed here are based on the insights of a previous pilot study [10], on achromatic HDR images, which inspired the work presented here; two measures (pixel-based DR and image key) are retained from the study and another three (area, contrast and colorfulness) are included based on analysis of that study's results and previous findings on these topics [17, 24, 11]. Three objective measures were used in the pilot study: dynamic range (DR), image key (IK) and spatial perceptual information (SI). After obtaining MOS scores, a significant correlation was found between the DR and MOS (Spearman's rank-order correlation coefficient $r_s = .788$; $p < .001$) and IK and MOS ($r_s = -.671$; $p < .001$), while the correlation between the SI and MOS was not found to be significant ($r_s = .037$; $p = .830$).

In addition, analysis of the results revealed that there might be a significant contribution to the perceived dynamic range based on the area of the brightest parts of the image. The images with relatively small light sources (e.g. *Zentrum, WaffleHouse, LasVegasStore*) had the highest difference between the perceived and pixel-based DR. This is in accordance with the discussions in the literature on image contrast and anchoring theories where the area and distance between the contrasting patches have been found to be the significant image features [32, 3]. In these studies, usually two types of *white* are discriminated: the diffuse, below a certain threshold; and specular or self-luminous, that represent values above the diffuse white threshold. A couple of studies on preferred diffuse range reported similar findings. The ITU document 6C/146-E, revealed that 90% of participants were satisfied with an upper limit of diffuse white being set to 2,400 cd/m^2 [33]. In a second study, the same threshold (90% preference level) was been reported for three groups of participants with the following values: 4,677 cd/m^2 for technical, 3,090 cd/m^2 for arts and 1,995 cd/m^2 for naive participants [34]. Both experiments were conducted on a dual modulation HDR display system [35], with the ability to reproduce luminance levels in the range from 0.004 to 20,000 cd/m^2. Seven images were evaluated, six of which were real-world structured stimuli, using 34 participants.

In addition, dynamic range is often confused with contrast in some contexts. Nevertheless, it was necessary to investigate perceptual relation between the two attributes, by inspecting whether a predictor of local contrast can further explain DR perception. Finally, as a number of psychophysical phenomena, such as the Helmhotz-Kohlrausch effect, link color appearance to dynamic range and contrast, colorfulness has been selected as an objective measure for chromatic images. The main goal was to investigate whether an attribute such as DR, generally considered monochromatic, could be perceived differently when the chrominance changes even if the luminance is kept constant.

3.2. Objective Metrics

Following the results of the pilot study, DR and IK were maintained as metrics and were augmented by another three measures based on the above observations - the *Area* of specular regions, the *Contrast* and the *Colorfulness*. The *Area* was motivated by the diffuse white thresholds; the *Contrast* as the magnitude of the global lightness difference is considered as a crucial factor in perception; and the *Colorfulness* since it is directly related to the perceived brightness of regions with constant luminance.

The pixel luminance values were first scaled to the display

range with the following equation:

$$L' = \frac{L - \min(L)}{\max(L) - \min(L)} \cdot (Disp_{\max} - Disp_{\min}) + Disp_{\min}, \quad (1)$$

where $Disp_{\min} = 0.03\,cd/m^2$ and $Disp_{\max} = 4250\,cd/m^2$ in our setup. The DR is then calculated after excluding 1% of the darkest and brightest pixels in the image, using

$$DR = \log_{10} \frac{\max(L')}{\min(L')}, \quad (2)$$

where L' is the image with scaled values.

The image key $IK \in [0, 1]$, a measure proposed by Aküz and Reinhard [36], was defined as:

$$IK = \frac{\ln(\text{avg}(L)) - \ln(\min(L))}{\ln(\max(L)) - \ln(\min(L))}, \quad (3)$$

where the avg(L) was computed as $\ln(\text{avg}(L)) = \Sigma_{ij} \ln(L(i, j) + \delta)/N$, with $\delta = 10^{-5}$ to avoid singularities and N was the number of pixels. Once again, min(L) and max(L) were calculated robustly, after excluding 1% of the darkest and the brightest pixels.

The *Area* was calculated as:

$$Area = \Sigma_{ij}(L(i, j)), L(i, j) > 2400\,cd/m^2, \quad (4)$$

and represents the number of pixels greater than the diffuse white threshold value. The value of 2,400 cd/m^2 was selected as recommended in ITU document [33], as discussed in Section 3.1. It is in accordance with the results of the study by Daly et al. [34], where different thresholds were found for different levels of user expertise. As participants in our study ranged from naive to technical, but not necessarily in this particular domain, it was reasonable to select a value towards to middle of the interval.

The *Contrast* measure is an adapted version of the work by Peli [20] and Rizzi et al. [21]. Let L^j be the image at level j of a Gaussian pyramid, of size $N/2^j \times M/2^j$, obtained by decimating the original image with a Gaussian low-pass filter, and let $N_8(L^j)$ be the 8-neighborhood for a given pixel in L^j. In this study, three levels were used, $j \in [2, 4]$, corresponding to 1.24, 2.47 and 4.95 cycle per degree (cpd), as the peaks of the contrast sensitivity function (CSF) for different light adaptation luminance lie approximately within that spatial frequency interval [37]. The cycles per degree were calculated based on the image dimensions, screen size and viewing distance. The contrast was computed as:

$$C = \frac{\sum_{j \in [2,4]} LC^j}{3}, \quad (5)$$

where

$$LC^j = \frac{\sum \frac{|L^j_{x,y} - N_8(L^j_{x,y})|}{8}}{L^j_W \times L^j_H}, \quad (6)$$

uses perceptually uniform luminance values [38]. L^j_W and L^j_H are the width and the height of the image L at the j-th scale.

3.3. Experimental Method

The main aim of this study was to investigate the correlation between PDR and the five objective measures, for both achromatic and chromatic images, and, subsequently, to propose a model for calculating dynamic range based on these measures. Below, the design, participants, apparatus, stimuli and the experimental procedure are described.

3.3.1. Design

In the study, a subjective evaluation of PDR of both achromatic and chromatic images was conducted. The participants were asked to evaluate the *overall impression of the difference between the brightest and the darkest part(s) in the images*. The independent variable was the image content, while the dependent variable was the reported PDR of the image. The study was conducted in two separate sessions, one for achromatic and another for chromatic images; at least one day was allowed between sessions. The sessions were ordered randomly and equally.

Three possible evaluation methods are typically considered during the design of experiments of this type: paired comparison, ranking and rating. Paired comparisons were ruled out due to their impracticality with large data sets. The efficient pair comparison techniques [39] can be used under certain assumptions. However, due to multidimensionality and non-deterministic DR appearance, these assumptions in our case were violated. While the ranking methods are straightforward, and quick to conduct, as with pairwise comparisons, they provide no information on the magnitude of the differences. Therefore, this method has been designed in order to use the advantages of the three methods: it permits ranking of the stimuli, a direct comparison between the image pairs, and it uses the continuous scale for subjective scores.

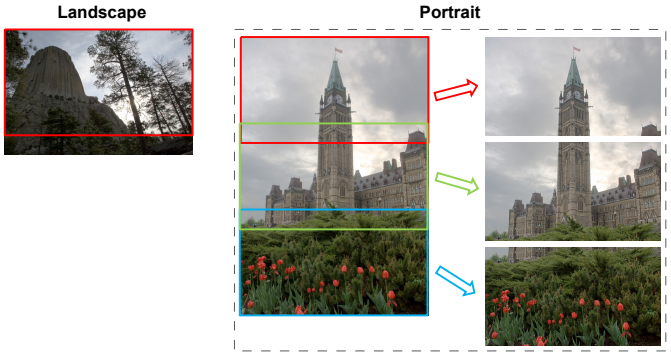


Figure 1: Preprocessing of the two images from the Fairchild's data set: the landscape image (left) was first downsized from 4288×2848 to 1920×1275, and then cropped to 1920×1080 pixels; the portrait image (right) was first downsized from 2704×4015 to 1920×2851 pixels and the three HD images were cropped - top, middle and bottom.

The evaluation method was inspired by the Subjective Assessment Methodology for Video Quality (SAMVIQ) [12], adapted to static images. The data set consisted of 36 images,

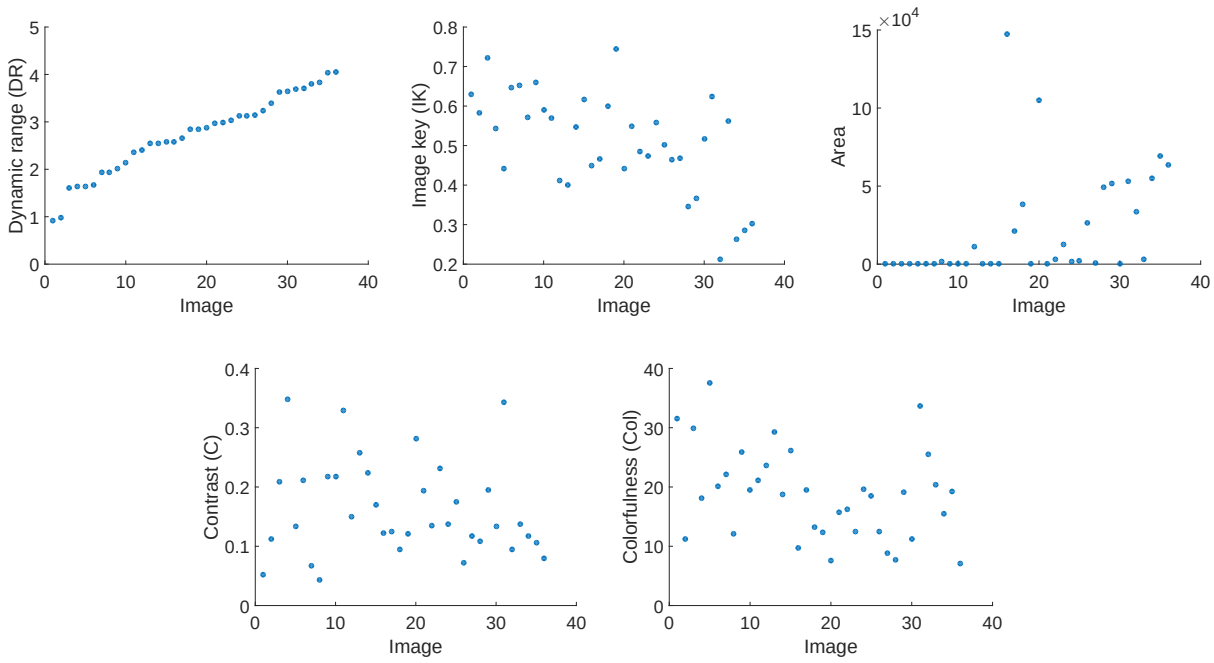


Figure 2: Image features: pixel-based DR (top-left), IK (top-middle), Area (top-right), Contrast (bottom-left) and Colorfulness (bottom-right), all sorted by pixel-based DR.

selected from the pool of 137 images, and divided into three subsets of 12 pseudo-randomly selected images in a randomized order (see section 3.3.4). The images thumbnails (422 × 238 px) were presented across a 3 × 4 grid, with the corresponding subjective scores, initially set to zero, below each image. The red color of the score indicated that the image had not been yet evaluated. All thumbnails were tone-mapped [40] for two main reasons: the images were rather small, thus making it inappropriate for subjective evaluation; and to discourage ratings based on the thumbnail appearance only.

The evaluation session was not time constrained. Each subset was evaluated independently, allowing participants to re-evaluate any image within, but not across subsets as many times as they wanted. The rating was performed on a 0-100 vertical continuous scale, divided into five equal intervals with corresponding labels: very low, low, medium, high and very high, included for general guidance. Upon completion, all participants were interviewed, using a short structured questionnaire as described in Section 3.3.5.

3.3.2. Participants

24 participants volunteered for the study. All of them reported normal or corrected-to-normal visual acuity and were screened for color blindness using the Ishihara test prior to the session with chromatic images. One male participant (age 49) was found to be color blind before proceeding to the second session with chromatic images, and therefore his scores for the achromatic stimuli were discarded and not used in the analysis. From the remaining 23 participants 17 were male and 6 female, with the age ranging from 23 to 40 (with an average age of 29). 12 participants were assigned to the chromatic condition for their first session, while the other 11 first evaluated the achromatic images during their first session.

3.3.3. Apparatus

All the experiments took place in a dark and quiet room. The stimuli were displayed at full HD (1920 × 1080 pixels) on an HDR SIM2 HDR47ES4MB 47" screen that allows for displaying >90% of Rec 709 color gamut [41]. It was utilized in the DVI Plus (DVI+) mode, that allows for directly and independently controlling backlight LEDs and LCD pixel values, based on the dual-modulation algorithm [42]. The ambient illumination in the room was measured between the screen and participants at 2.154 *lux*. The luminance of the screen when turned off was 0.03 cd/m^2. The distance from the screen was fixed to three heights of the display, with the eyes in the middle of the display, both horizontally and vertically.

3.3.4. Stimuli

Initially, all the images from the HDR Photographic Survey [32] were considered. The landscape images were downsized and cropped to 1920 × 1080 size, while all the portrait images were first downsized to 1920 pixel width and then three images were cropped out: the top-aligned, the middle-aligned and the bottom-aligned one, see Figure 1. This resulted in a total of 131 landscape HD HDR images. After computing the DR, IK and SI [10], a set of 33 images was selected for the study; selected to maintain objective measures evenly distributed across the set. The features used in this study show the same trend, see Figure 2. Since most of the images from the Fairchild's data

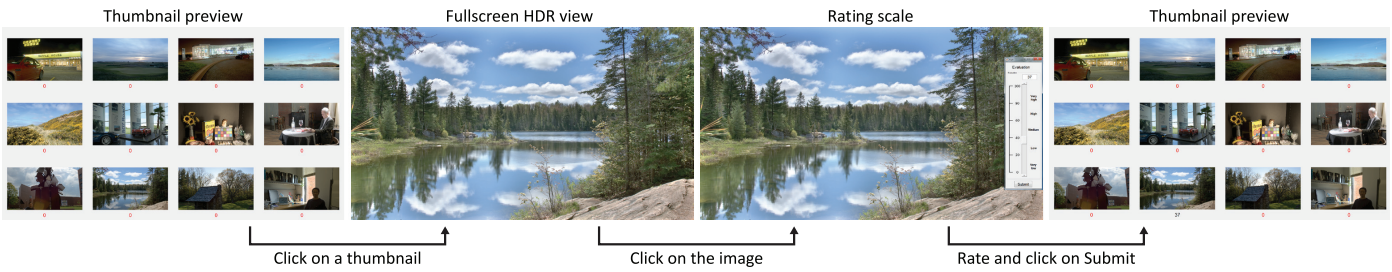


Figure 3: The illustration of the test procedure.

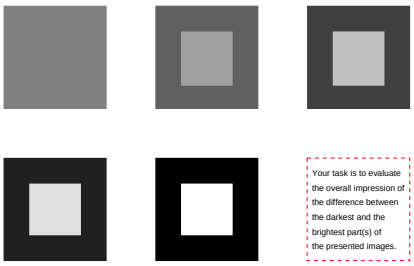


Figure 4: The abstraction of the attribute to be evaluated.

set are photographs of nature, a single frame from the *Market* HDR video sequence proposed in MPEG by Technicolor [43] and a frame from both the *Carousel* and *Bistro* video sequences from the Stuttgart HDR Video Database [44] were added to the test data set, see Figure 5. All 36 images were converted to the corresponding achromatic images, using BT.709 primaries to compute relative luminance [45]. In order to reproduce an image on the display, the luminance values should not exceed the range of the display. Values in excess of the maximum display brightness, for both the chromatic and achromatic images, were first clipped to the diplay's peak luminance value of $4250 cd/m^2$, and the images were then processed using the algorithm developed by Zerman et al. [42] and displayed as RGB images in DVI+ mode.

From the set of 36 test images, six images with the lowest and six with the highest pixel-based DR were selected in two new subsets: *imagesLow* and *imagesHigh*. When generating each session for the experiments, each subset of 12 images was composed of two randomly selected images from the *imagesLow* subset, two from *imagesHigh* and the rest from the remaining 24 images. This preserved the consistency of measures among the subsets.

3.3.5. Procedure

Upon entering the experimentation room, the participants were first given the instructions to read and asked if they had any questions about the nature of the experiment and their task. This was followed by a training session, where the task was first explained using feature abstractions, inspired by the study by Aydin et al. [8], see Figure 4. They were told that in this example the magnitude of the feature increases from minimum in the first gray block to maximum in the last, black and white block. After this, the simulation of the experimental framework was displayed with three images with high, low and average pixel-based DR range on each screen. The first three images were evaluated by the experimenter explaining that they correspond to the top, bottom and the middle of the rating scale respectively, and the rest of the images by the participant in order so stabilize their opinion. None of the six training images were part of the test data set. In both the written and verbal instructions, they were asked to evaluate the overall impression of the difference between the brightest and the darkest part(s) of the images. If the participants demonstrated a full understanding of the experimental task, they were asked to fill in the basic information form and the corresponding session commenced.

The images in each subset were evaluated independently; once the participant selected the next subset the results of the previous one could not be changed. However, they could evaluate all the images in the current subset in any order and as many times as they needed. This allowed for multiple comparisons between the images and fine adjustments of the scores. In order to evaluate an image, the participant had to click on the thumbnail. The selected HDR image was displayed full screen for evaluation. Once the participant was ready to give a rating they clicked on the presented image and a rating scale was produced. The scale appeared in the far right side of the screen. After the score was given, the initial thumbnail preview of the current subset was re-displayed with the updated score for the evaluated image, see Figure 3. After the completion of the test, the experimenter had a short structured discussion about the test with all the participants. The questions used in the experiment were:

1. On a scale 1-10, how tiring was the experiment in terms of visual comfort and fatigue?
    - Have you been bothered by some particular images?
2. On a scale 1-10, how difficult did you find it to evaluate the images?
    - What did you find difficult? Why? Which (type of) images?
3. With how many ratings are you confident (you think your score is correct)?
4. Would you change anything in the experiment?

507


AirBellowsGap


BandonSunset(1)


BigfootPass


Bistro


BloomingGorse(1)


Carousel


DelicateFlowers


DevilsTower


ElCapitan_b

Flamingo


GoldenGate(1)


HancockKitchenInside


HancockKitchenOutside


HDRMark


JesseBrownsCabin

LabBooth


LabTypewriter

LasVegasStore


Market3


NorthBubble


OCanadaNoLights_b


OCanadaNoLights_m


OtterPoint

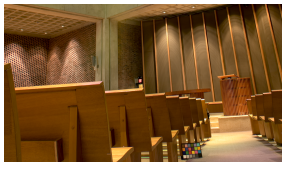

PaulBunyan


PeckLake


SequoiaRemains_t

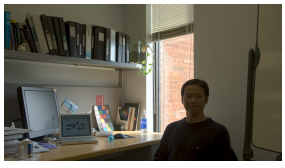

TupperLake(1)


URChapel(1)_t

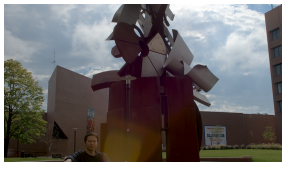

URChapel(2)_b


URChapel(2)_m

WaffleHouse


WillyDesk


WillySentinel_b


WillySentinel_m


Zentrum

Figure 5: Thumbnails of the images used in the study.

Table 2: Pearson's r and Spearman's r_s correlation coefficients between MOS values for both achromatic and chromatic images and five objective measures: DR, IK, Area, C and Col. * denotes significance at p < .01.

MOS	Achromatic				Chromatic				
Measure	DR	IK	Area'	C	DR	IK	Area'	C	Col
r	.87*	-.60*	.87*	-.19	.84*	-.61*	.87*	-.22	-.47*
r_s	.87*	-.55*	.89*	-.24	.84*	-.57*	.90*	-.27	-.43*

- If so, what? How? Why?

5. Is there anything else you would like to comment on?
6. Was it easier to evaluate grayscale or color images? [2]

4. Results

Results were analyzed with a series of statistical tests to form a better understanding of the captured data. This section presents results across the participants testing for group, image and chromatic conditions and correlations comparing MOSs with the objective measurements.

4.1. Overall Results Across Participants

The effect of the independent variables on the PDR scores provided by the participant was analyzed in a 2 (*session*) × 2 (*color*) × 36 (*scenes*) factorial design. *session* is a between-participants variable reflecting the order of chromatic-achromatic presentation whereby 12 participants were first presented with the chromatic condition and the other 11 with the achromatic condition on their first session. *color* is a within-participants variable that corresponds to the chromatic and achromatic results and *scenes* is also a within-participants variable corresponding to the 36 images presented to the participants.

Results were analyzed using a mixed-model factorial ANOVA. The main effect of *session* was not significant F(1, 21) = 2.06 , p = .17. This indicates that the session ordering did not have a significant effect on the results.

The main effect of *color* was also insignificant, F(1, 21) = .40 , p = .40. Again this indicates that there was no significant difference between the scores for the chromatic and achromatic set. Further analysis of the chromatic and achromatic results is given below.

The main effect of *scenes* was significant F(35, 735) = 98.73 , p < .01 as expected. Analysis of the differences and correlations with objective measurements will be given below.

In order to analyze the agreement across participant scores, Kendall's coefficient of concordance W was employed. W reports a value of 1 for absolute agreement across all judges and 0 for complete disagreement. For the achromatic condition, Kendall's coefficient of concordance was significant (W = .81, p < .01) and similarly for the chromatic condition (W = .76, p < .01). These results indicate a very strong agreement across participants' allocated scores.

4.2. Correlations of MOS

In order to analyze the correlations among the PDR values provided by the participants per scene, results were collapsed into a mean opinion score (MOS) per scene. The MOSs are correlated with the four (five for chromatic) objective measures presented in Section 3.2. The scatter plots for each of the five objective measures are provided in Figure 6 and Figure 7.

Since the calculated measures were on different scales, they were scaled using the equation:

$$x_i' = \frac{x_i - \frac{1}{n}\sum_{i=1}^{n} x_i}{max(X) - min(X)} \tag{7}$$

so that they are all represented with the same order of magnitude. The results of the correlation between the MOS values and objective measures can be seen in Table 2.

Initially, when comparing the Pearson's (r) and Spearman's (r_s) correlation coefficients, the only substantial difference was found for the Area measure (r = .64 and r_s = .87 for achromatic, and r = .64 and r_s = .88 for chromatic stimuli). By observing the correlation scatter plots (Figures 6 and 7), an evident trend of the root function was noticeable for this measure. A number of root functions were tested in order to linearize the data and the best was found to be:

$$Area' = \sqrt[4]{Area}. \tag{8}$$

With the fitted data ($Area'$) the Pearson's correlation coefficient was almost identical to the Spearman's (see Table 2), which eventually resulted in a more robust predictor model, see Section 5.

The correlations between the chromatic MOS and achromatic MOS were r = .99 and r_s = .98, both significant at p < .01, indicating the chromatic and achromatic scores were very highly correlated.

In Figure 8, the extended box plot depicts the distribution of the perceived DR scores, with the corresponding mean and median values, confidence intervals and outliers. In addition, the pixel-based DR is also presented to visually display the correlation between the subjective and objective scores.

4.3. Post-experimental Inquiry

The comments obtained upon completion of each session of the study were fairly consistent across all the questions. The average score of the answer to the first question was 3.93, which means that there were instances which were slightly annoying

[2] This question was asked only upon completion of the second session.

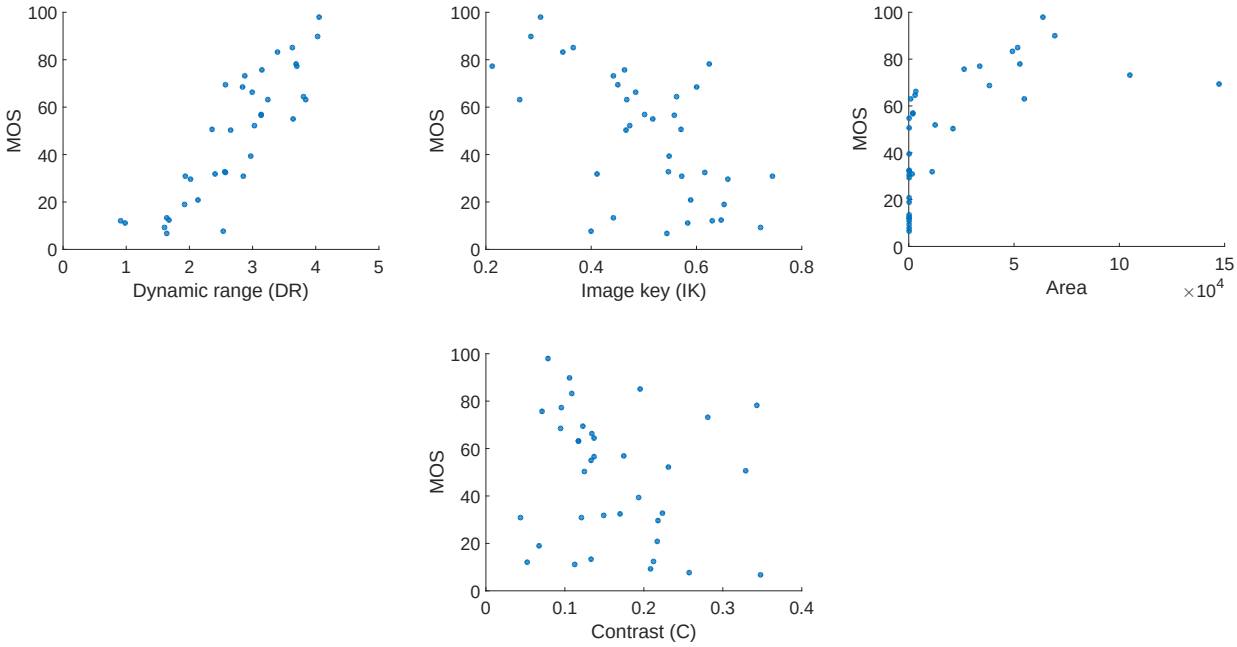


Figure 6: Achromatic MOS across the four objective measures (top-left to bottom-right): pixel-based DR, IK, Area and Contrast.

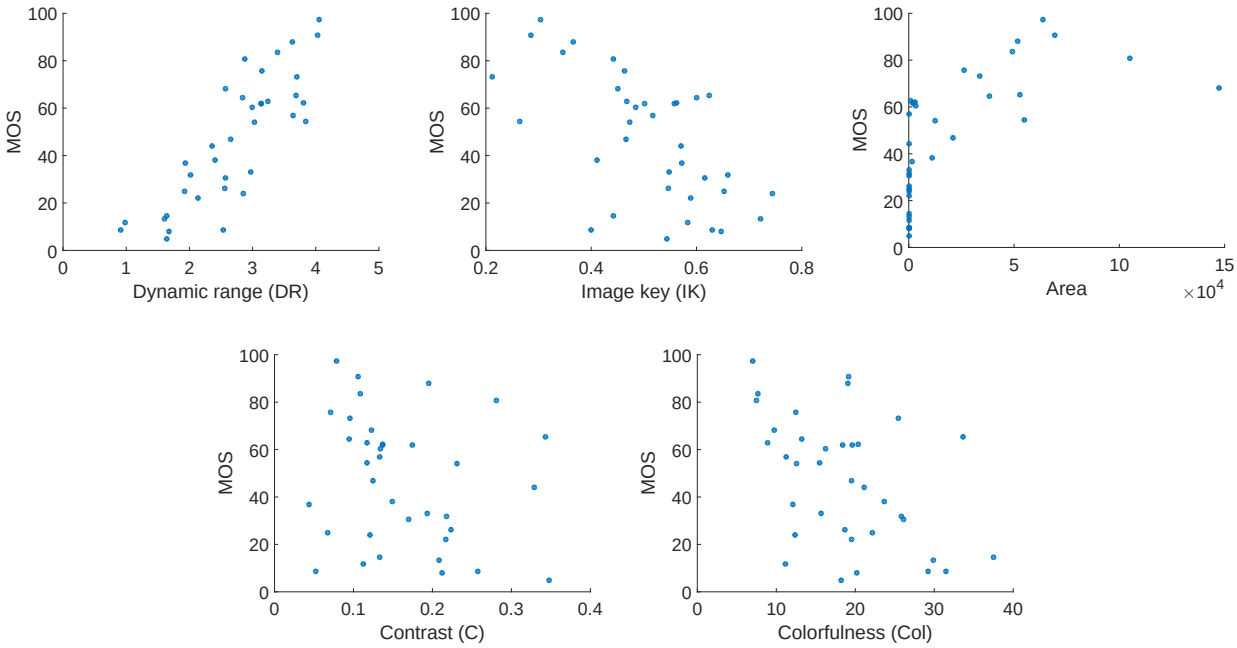


Figure 7: Chromatic MOS across the five objective measures (top-left to bottom-right): pixel-based DR, IK, Area, Contrast, Colorfulness.

to observe. Around 60% of participants reported that the brightest images with the visible sun or light sources were uncomfortable to look at.

The average score to the second question was 4.17. Several participants reported that the images with "very small but very bright portions of the image" were more difficult to evaluate and that the area of the brightest parts (i.e. light sources, sun) could have an affect on the perception of DR. A number of participants mentioned that they think that the PDR varies depending on the proximity of the brightest and the darkest parts in the image. For example, if there is the sun in one corner of the image and a part of the scene in shadow in the other corner (e.g. *OtterPoint*) versus if there is the sun behind the tree (e.g. *DevilsTower*).

It was reported that the user confidence for the given scores (question 3; i.e. images that were relatively easy to evaluate and to which they think they gave an objective score) was 70.74%. The images constituting the remaining 29.26% correspond to the images with a medium DR, which was in accordance with the increased confidence interval for these images, see Figure 8.

Generally, all the participants liked the experimental design and there were no major complaints regarding the framework (question 4). One participant suggested that it might be better to have more than 12 images in a subset. Initially, there were two subsets, each consisting of 20 images. However, the pilot study revealed that it might be too difficult to perform comparisons among that many images. Furthermore, ten images are used in the SAMVIQ methodology, upon which this framework had been developed.

In the general comments section (question 5), the participants reported that the possibility of re-evaluating images was very helpful and liked the fact that they could evaluate images in any order without time constraints. They also reported that the overall image brightness might be affecting the PDR.

The responses to the last question revealed that it was slightly easier to evaluate achromatic images (60.87% of participants) as opposed to chromatic (26.09%). 13.04% of participants reported that it was the same in terms of difficulty. Most of the participants from the first group (i.e. the 60.87% who

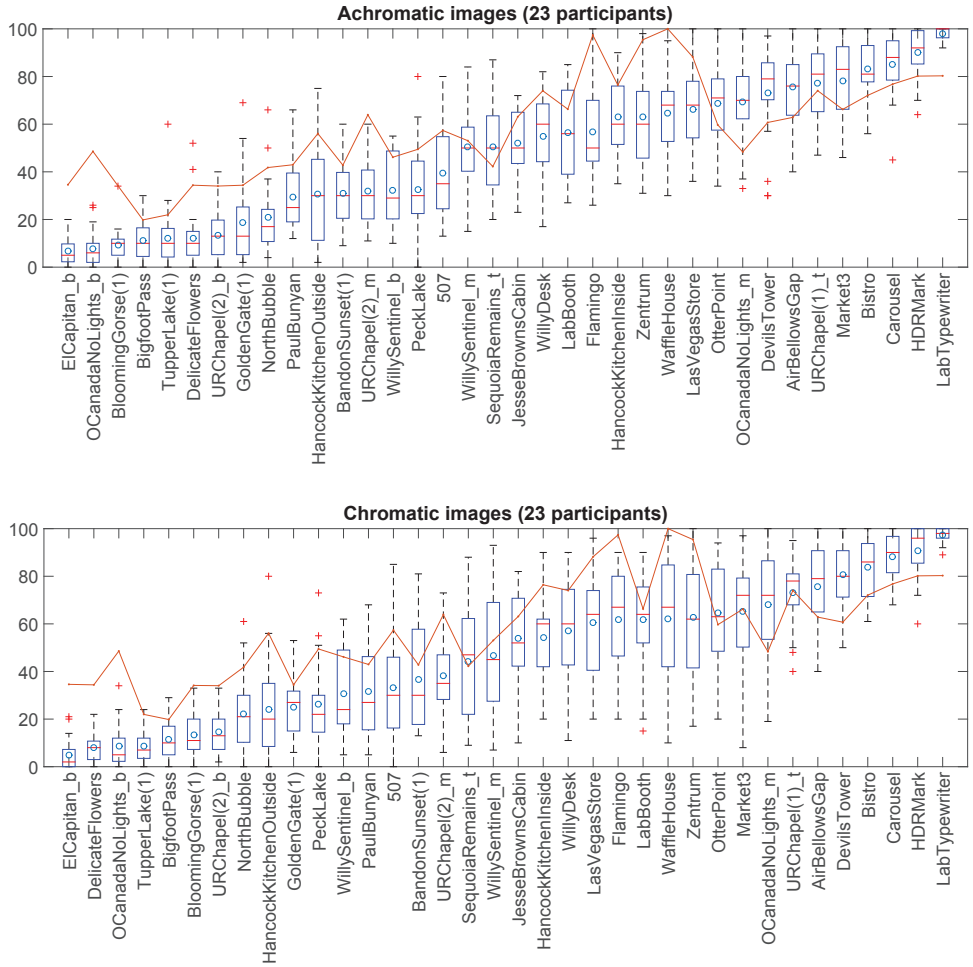


Figure 8: Extended boxplot diagrams for achromatic (top) and chromatic (bottom) images. Blue circles = MOS; Red horizontal lines = median values; Blue boxes = the interquartile ranges; Whiskers = adjacent values; Red crosses = outliers; Red line: pixel-based DR values (scaled as $DR = \frac{DR}{\max(DR)} \cdot 100$). The scores are sorted by the mean value.

Table 3: The summary of the model with R, R^2, adjusted R^2 and F change significance values for both achromatic and chromatic images, presented across the hierarchies.

	Achromatic				Chromatic			
Model	R	R^2	adj. R^2	sig. F change	R	R^2	adj. R^2	sig. F change
DR	.866	.750	.743	.000	.839	.704	.695	.000
DR, Area'	.945	.893	.887	.000	.932	.868	.860	.000

found achromatic easier) reported that this could be due to the extra chromatic information which compounded the complexity of the choice.

These comments were a valuable input to the study and were considered when choosing the predictors in the regression model (see Section 5).

5. PDR Predictor Model

This section looks into producing a model that can predict PDR based on the results of the experiment. To achieve this, multivariate linear regression is employed and analyzed.

5.1. Creating the Model

The results from the pilot study [10] and the previous section indicate that the pixel-based DR is a generally good predictor of the PDR, but there are cases where it fails (see Section 6), and that it could be potentially improved if other measures are considered. The results were also used as an indication of the measures that could be used as variables for predicting the outcome in the multiple regression analysis. In particular, a model is fit to the data and used to predict values of the outcome from several predictors.

Miles and Shelvin [46] suggest that when expecting a large effect size, a sample size of around 40 is sufficient for two to four predictors. This is close to our initial sample of 36 images. Prior to employing the regression, the data for the Area measure was linearized using the Equation 8, and all the measures were feature-scaled using Equation 7. The hierarchical method was utilized in order to see the behavior of the model with new predictors. Since pixel-based DR was known to be a good predictor of the perceived DR, it was selected for the first block in the hierarchy. The forced entry method was selected for DR. All other predictors (IK, Area', C and Col) were added to the second block. For this block, the forward method was used. This method calculates the contribution of each predictor by looking at the semi-partial correlation with the outcome. The model summary with the corresponding R, R^2, adjusted R^2 and the significance of the F change values are presented in Table 3.

The R values show the multiple correlation coefficients between the predictors and the outcome. In the first case where there is only one predictor (DR), this is a simple correlation between MOS and this measure. In case of achromatic images, the R value increases from .866 with DR only, to .945 with DR and Area' measures. Similar R values can be observed for the chromatic images (R = .839 for DR and R = .932 for DR and Area' measures).

Table 4: Estimated *BetaIn* value, t-statistics and its significance for each predictor. For both achromatic and chromatic models the Area' predictor has the highest (significant) t value, and was therefore included in the model. In the following iteration none of the remaining variables (IK, C and Col for the chromatic model) had significant t value, and thus were not included into the model.

	Achromatic			Chromatic		
Model	*BetaIn*	t	$Sig.$	*BetaIn*	t	$Sig.$
IK	-.137	-1.303	.202	-.181	-1.610	.117
Area'	.516	6.657	.000	.553	6.418	.000
C	-.113	-1.325	.194	-.150	-1.640	.110
Col				-.219	-2.362	.024
IK	.069	.889	.381	.031	.361	.720
C	-.049	-.838	.408	-.082	-1.291	.206
Col				-.116	-1.725	.094

Table 5: Model parameters for the achromatic and chromatic images.

	Model	*B*	*Std. Error*	*Beta*	*t*	*Sig.*	*VIF*
Achromatic	Constant	$6.221e^{-6}$	.017		.000	1.000	
	DR	.573	.086	.515	6.647	.000	1.859
	Area'	.448	.067	.516	6.657	.000	1.859
Chromatic	**Model**	*B*	*Std. Error*	*Beta*	*t*	*Sig.*	*VIF*
	Constant	$-1.795e^{-6}$	.018		.000	1.000	
	DR	.506	.094	.463	5.383	.000	1.859
	Area'	.471	.073	.553	6.418	.000	1.859

In the stepwise regression, at each iteration, the *BetaIn* value is estimated for each predictor that has not been included into the model, as if it were entered into equation at this stage. The standardized b-value, *BetaIn*, indicates the number of standard deviations that the outcome will change as a result of one standard deviation change in the predictor. Based on this value, the t-statistics for these values are computed and based on its significance the next predictor is entered, until there are no predictors with significant value less than .05, see Table 4.

5.2. The Model Parameters

After finding the independent variables that significantly improve the prediction of the outcome variable, the model coefficients, for the model using DR and Area' were calculated, see Table 5.

Therefore, the regression model for the achromatic images is defined as:

$$MDR \approx 0.573DR + 0.448\sqrt[4]{Area} \tag{9}$$

while for the chromatic images it is:

$$MDR \approx 0.506DR + 0.471\sqrt[4]{Area} \tag{10}$$

In Figure 9 the observed scores (MOS), pixel-based DR and MDR values are displayed for each scene. A significantly higher correlation between the MOS and MDR than between the MOS and pixel-based DR is evident. The cases with the highest discrepancy between the latter will be further discussed in Section 6.

In order to verify whether the model fits the data or if it is skewed by a few extreme cases (outliers), the standardized residuals (SR), obtained by dividing the residual by an estimate of their standard deviation, were examined. In addition, Cook's distance (CD) was calculated in order to check for the influential cases [47]. This test analyzes whether the regression model is stable across the sample, by calculating the overall influence of a particular case on the model. Based on the findings by Cook and Weisberg [48], values greater than 1 may be cause for concern. Computing the residual statistics on a case-wise basis, the results revealed that there were no cases with the absolute value of the SR greater than 3.29. For three achromatic images: *HancockKitchenInside*, *OCanadaNoLights_bottom* and *SequoiaRemains_top* this value was −2.25, −2.441 and 2.238 respectively. In the case of chromatic images there was only *HancockKitchenInside* image for which the SR value was −2.748. Cook's distance (CD) for the three achromatic images was .170, .167 and .087 respectively, while for the *HancockKitchenInside* chromatic image this value was .253 respectively. Therefore, based on these results, the sample represents an accurate model.

Multicollinearity between predictors makes it difficult to assess the individual contribution of a predictor. If the two pre-

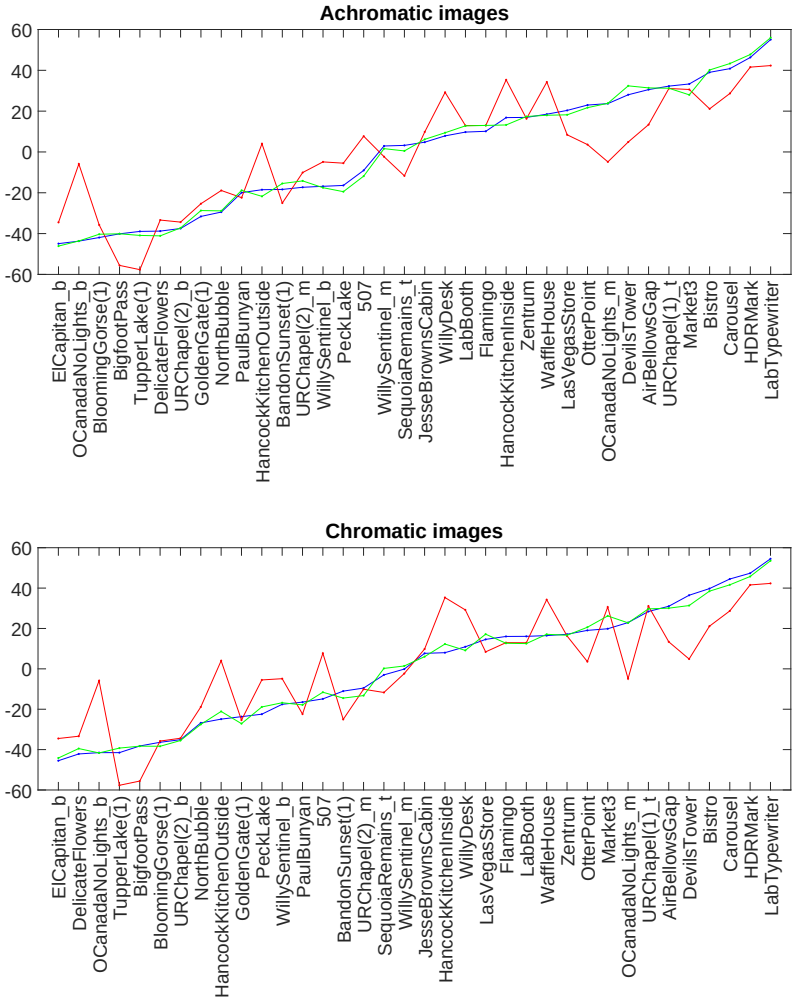


Figure 9: Graphical representation of the MOS (blue), pixel-based DR (red) and MDR (green) values for achromatic (top) and chromatic (bottom) images on feature-scaled values computed using Equation 7. There is a high correlation between the MOS and MDR, and significant deviations in some cases of traditional DR computation. The scores are sorted by the MOS value.

dictors are highly correlated, there is a high possibility that each accounts for similar variance in the outcome. Multicollinearity was tested by using the variance inflation factor (VIF), see Table 5. As argued by Myers, a value greater than 10 indicates a potential problem [49]. In addition, values less than 0.2 indicate a potential multicollinearity and should be further investigated [50]. As VIF values for both achromatic and chromatic models were 1.859, there is no indication of multicollinearity between the predictors in the models.

5.3. Calculating Model Generalizability

The R^2 values show how much of the variability in the MOS is accounted for by the predictors. For the achromatic images using only DR as a predictor $R^2 = .75$, while for the same model with chromatic images $R^2 = .704$, which means that DR accounts for 75% and 70.4% of total variation of the MOS respectively. Adding Area' as the second predictor, this percentage rises to 89.3% for achromatic, and 86.8% for chromatic.

Finally, for assessing how well the sample represents the entire population, that is how accurately the model can predict the outcome in a different sample, a cross-validation of the model was performed. If the prediction on another sample is similarly correct, then the model can be generalized. Cross-validation was calculated by adjusting the R^2 values by estimating how the R^2 values were derived from the population from which the sample was taken. The adjusted R^2 values give an indication of how well the model generalizes. The closer the values to the R^2 ones are, the better the prediction from the sample is. For example, the difference between the R^2 and the adjusted R^2 values for the achromatic model with two predictors is $0.893 - 0.887 = 0.006$, which means that if the model was derived from the population, instead of from the sample, it would account for 0.6% less variance in the MOS. The values presented in Table 3 indicate a very good cross-validity of the model.

Calculating the change statistics, the significance of the change in R^2 by adding new predictors can be calculated. This is usually done by calculating the F-ratio by using the following equation:

$$F = \frac{(N - k - 1)R^2}{k(1 - R^2)} \tag{11}$$

where N is the number of cases, and k is the number of predictors in the model.

In Table 3, a significance of the F change is provided. In both the achromatic and chromatic image models, there was a significant change in R^2 value for addition of the Area' ($p < .001$).

6. Analysis of High Discrepancy Scenes

As was shown in Section 4, a straightforward metric such as pixel-based DR can be a good predictor of dynamic range for many of the images. However, there are some particular cases where it fails to truthfully predict the PDR. In this section the scenes with highest discrepancies between pixel-based DR and MOS values are investigated, for both the achromatic and chromatic images. Two subsets of eight scenes with the highest discrepancies were selected, see Figure 10. The values obtained by taking the absolute value of the subtracted feature-scaled scores are presented in Table 6.

Table 6: Eight achromatic (top) and eight chromatic (bottom) scenes with the highest difference between the MOS and pixel-based DR, along with the differences between the MOS scores and predicted perceived dynamic range by the proposed model (MDR). The scores are calculated as the absolute difference between the feature-scaled values obtained using Equation 7), and multiplied by 100 for better readability.

	Scene	**|MOS-DR|**	**|MOS-MDR|**
Achromatic	DevilsTower	23.148	**0.136**
	HancockKitchenInside	18.495	22.350
	HancockKitchenOutside	22.515	**9.251**
	OCanadaNoLights_b	37.795	**24.263**
	OCanadaNoLights_m	28.548	**2.277**
	OtterPoint	19.377	**4.977**
	TupperLake(1)	18.745	**10.167**
	WillyDesk	21.325	**1.539**
Chromatic	507	22.590	**5.074**
	Bistro	18.622	**10.155**
	DevilsTower	31.579	**7.605**
	HancockKitchenInside	27.317	29.778
	HancockKitchenOutside	28.914	**14.788**
	OCanadaNoLights_b	35.600	**21.639**
	OCanadaNoLights_m	27.772	**4.854**
	WillyDesk	18.280	**3.838**

These results demonstrate that in 14 out of the 16 cases the model predicts the PDR better than the pixel-based DR measure. Table 7 shows the correlations between the MOS scores and both pixel-based DR and MDR, performed on these subsets. For both achromatic and chromatic scenes, the correlation between the MOS and DR is not significant ($p > 0.05$). Nevertheless, the correlation between the MOS and predicted perceived dynamic range (MDR) is significant in all cases. These chosen cases demonstrate why the predictive method can be considered superior to the DR method when predicting PDR.

Table 7: The Pearson's r and Spearman's r_s correlation coefficients between MOS values and both pixel-based DR and MDR for achromatic and chromatic scenes. *Correlation significant at p < .01; **Correlation significant at p < .05.

MOS	Achromatic		Chromatic	
Measure	DR	MDR	DR	MDR
r	.584	**.912***	.354	**.881***
r_s	.429	**.833****	.357	**.762****

Figure 10: Achromatic (top) and chromatic (bottom) high discrepancy images used in the analysis.

7. Conclusions and Future Work

While traditional dynamic range measures can be an acceptable predictor of PDR, they can also be inaccurate in some cases. In order to develop a generic model that can truthfully predict PDR, sensory and cognitive processes involved in extraction of this attribute from complex stimuli need to be taken into account. Previous research has shown that the area of the brightest patches and the image topology affect the perception of lightness and contrast [17, 19, 26], and therefore should be considered in constructing such a PDR model. Furthermore, image contrast and colorfulness are important factors in visual perception and, based on previous findings [8, 11, 24, 25], could be involved in the process related to the extraction of observed image attributes.

In this study a new experiment for subjective evaluation of perceived dynamic range in HDR images has been designed and conducted. The results were used to generate a subjectively annotated data set of 36 HDR images, with both MOS scores for achromatic and chromatic data. This data set can be used in future HDR content studies on algorithm or metric validation and aesthetic attribute analysis and modeling. This is, to the best of the authors' knowledge, the first study on perceived DR using complex stimuli and HDR conditions.

The results of the Kendall's coefficient of concordance show that there was a high agreement between participants' scores in both sessions. Using the mixed model factorial ANOVA the effect of scenes was found as significant, while neither the session order nor chromaticity of the data were found to produce significant differences in the results. In addition, the correlation between the chromatic and achromatic MOSs was computed, and the results demonstrated very high correlations. Furthermore, the correlations between the PDR and five objective measures were computed. All correlations, except for the contrast (C), were found as significant for both achromatic and chromatic images, with very small discrepancies between the two conditions.

Models that can predict PDR of both achromatic and chromatic images were generated using multivariate linear regression. The regression model revealed that only pixel-based DR and linearized Area measures had significant contribution to the predictor model. Cross-validation of the model showed that the model could accurately predict the outcome in a different sample, i.e. the model can be generalized.

Finally, eight scenes with the highest discrepancies between the MOS and pixel-based DR values were selected from both achromatic and chromatic sets. The statistics showed that the PDR prediction was significantly improved when the Area predictor contributed to the model.

Although the results show that, in the overall, the PDR prediction with the proposed model is closer to the MOS it is likely that there will be images where this is not the case due to the excessive complexity of the HVS and the related processes in

the perception of such visual attributes. Therefore, in the future, we would like to further investigate this topic by looking at these and other perceptual factors that could be involved in this process at a lower level. Furthermore, the existing objective metrics have to be redesigned and, possibly, novel ones developed targeted directly at HDR content. While this is beyond the scope of this paper, it was evident that a gap exists in this area. Another direction at which we would like to expand this work is the analysis of aesthetic attributes. Finally, we are interested in extending this work to video content and investigating the temporal aspects. In the case of video, the perceptual phenomena behind the perception of dynamic range can be more complex. While, for a static image, the luminance range reproducible by an HDR display matches the steady-state dynamic range of the HVS, temporal variations of this range, e.g., due to a change from a bright to a dark scene, can span a much broader interval of luminance than the HVS could process at a given adaptation level, causing maladaptation phenomena and visual discomfort [37]. It is known that light/dark adaptation is not instantaneous, which results in higher masking for larger temporal variations of the luminance range. This entails a loss of contrast sensitivity in the maladaptation phase, but could enhance the overall perception of bright-dark differences on short time segments. Therefore, initial studies will be conducted with a similar methodology, using short clips, and the scores will be correlated with dynamic range models similar to those discussed in this work.

8. Acknowledgments

The authors would like to thank Emin Zerman for the help with the DVI+ content generation and with development of the test framework. We would like to thank all the participants that volunteered in the subjective study. This work was supported by Region Ile de France, in the framework of the FUI 4EVER2 project. Debattista is partially supported by a Royal Society Industrial Fellowship.

References

[1] F. Banterle, A. Artusi, K. Debattista, A. Chalmers, Advanced high dynamic range imaging: theory and practice, CRC Press, 2011.

[2] F. Dufaux, P. Le Callet, R. Mantiuk, M. Mrak, High Dynamic Range Video. From Acquisition, to Display and Applications, Academic Press, 2016.

[3] A. Gilchrist, C. Kossyfidis, F. Bonato, T. Agostini, J. Cataliotti, X. Li, B. Spehar, V. Annan, E. Economou, An anchoring theory of lightness perception., Psychological review 106 (4) (1999) 795.

[4] F. Banterle, P. Ledda, K. Debattista, A. Chalmers, Inverse tone mapping, in: Proc. of the 4th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia, GRAPHITE '06, ACM, New York, NY, USA, 2006, pp. 349–356. doi:10.1145/1174429.1174489.
URL http://doi.acm.org/10.1145/1174429.1174489

[5] A. G. Rempel, M. Trentacoste, H. Seetzen, H. D. Young, W. Heidrich, L. Whitehead, G. Ward, LDR2HDR: On-the-fly reverse tone mapping of legacy video and photographs, in: ACM SIGGRAPH 2007 Papers, SIGGRAPH '07, ACM, New York, NY, USA, 2007. doi:10.1145/1275808.1276426.
URL http://doi.acm.org/10.1145/1275808.1276426

[6] F. De Simone, G. Valenzise, P. Lauga, F. Dufaux, F. Banterle, Dynamic range expansion of video sequences: A subjective quality assessment study, in: IEEE Global Conference on Signal and Information Processing, IEEE, 2014, pp. 1063–1067.

[7] C. Mantel, J. Korhonen, S. Forchhammer, J. Pedersen, S. Bech, Subjective quality of videos displayed with local backlight dimming at different peak white and ambient light levels, in: 7th Int. Work. on Quality of Multimedia Experience, IEEE, 2015, pp. 1–6.

[8] T. O. Aydin, A. Smolic, M. Gross, Automated aesthetic analysis of photographic images, Visualization and Computer Graphics, IEEE Transactions on 21 (1) (2015) 31–42.

[9] M. Narwaria, C. Mantel, M. Perreira de Silva, P. Le Callet, An objective method for High Dynamic Range source content selection, in: 6th Int. Work. on Quality of Multimedia Experience, 2014, pp. 13–18.

[10] V. Hulusic, G. Valenzise, E. Provenzi, K. Debattista, F. Dufaux, Perceived dynamic range of HDR images, in: Proceedings 8th International Workshop on Quality of Multimedia Experience (QoMEX2016), IEEE, 2016.

[11] M. D. Fairchild, Color appearance models, John Wiley & Sons, 2013.

[12] J.-L. Blin, SAMVIQ - Subjective assessment methodology for video quality, Rapport technique BPN 56 (2003) 24.

[13] S. S. BERGSTROM, Common and relative components of reflected light as information about the illumination, colour, and three-dimensional form of objects, Scandinavian journal of psychology 18 (1) (1977) 180–186.

[14] H. Barrow, J. Tenenbaum, Recovering intrinsic scene characteristics, Comput. Vis. Syst., A Hanson & E. Riseman (Eds.) (1978) 3–26.

[15] A. L. Gilchrist, The perception of surface blacks and whites, WH Freeman, 1979.

[16] E. H. Adelson, A. P. Pentland, The perception of shading and reflectance, Perception as Bayesian inference (1996) 409–423.

[17] X. Li, A. L. Gilchrist, Relative area and relative luminance combine to anchor surface lightness values, Perception & Psychophysics 61 (5) (1999) 771–785.

[18] E. H. Land, J. J. McCann, Lightness and retinex theory, JOSA 61 (1) (1971) 1–11.

[19] E. Provenzi, L. De Carli, A. Rizzi, D. Marini, Mathematical definition and analysis of the retinex algorithm, Journal of the Optical Society of America A 22 (12) (2005) 2613–2621.

[20] E. Peli, Contrast in complex images, JOSA A 7 (10) (1990) 2032–2040.

[21] A. Rizzi, T. Algeri, G. Medeghini, D. Marini, A proposal for contrast measure in digital images, in: Conference on Colour in Graphics, Imaging, and Vision, Vol. 2004, Society for Imaging Science and Technology, 2004, pp. 187–192.

[22] K. Matkovic, L. Neumann, A. Neumann, T. Psik, W. Purgathofer, Global contrast factor-a new approach to image contrast., Computational Aesthetics 2005 (2005) 159–168.

[23] A. M. Haun, E. Peli, 24.1: Measuring the perceived contrast of natural images, in: SID Symposium Digest of Technical Papers, Vol. 42, Wiley Online Library, 2011, pp. 302–304.

[24] A. J. Calabria, M. D. Fairchild, Perceived image contrast and observer preference i. the effects of lightness, chroma, and sharpness manipulations on contrast perception, Journal of imaging Science and Technology 47 (6) (2003) 479–493.

[25] A. J. Calabria, M. D. Fairchild, Perceived image contrast and observer preference ii. empirical modeling of perceived image contrast and observer preference data, Journal of Imaging Science and Technology 47 (6) (2003) 494–508.

[26] P. Vangorp, K. Myszkowski, E. W. Graf, R. K. Mantiuk, A model of local adaptation, ACM Transactions on Graphics (TOG) 34 (6) (2015) 166.

[27] S. Winkler, Chapter 5: Perceptual video quality metrics–a review, Digital Video Image Quality and Perceptual Coding (2006) 155–179.

[28] J. Kopf, W. Kienzle, S. Drucker, S. B. Kang, Quality prediction for image completion, ACM Trans. Graph. 31 (6) (2012) 131:1–131:8. doi:10.1145/2366145.2366150.
URL http://doi.acm.org/10.1145/2366145.2366150

[29] A. O. Akyüz, R. Fleming, B. E. Riecke, E. Reinhard, H. H. Bülthoff, Do HDR displays support ldr content?: a psychophysical evaluation, in: ACM Transactions on Graphics (TOG), Vol. 26, ACM, 2007, p. 38.

[30] F. Banterle, P. Ledda, K. Debattista, M. Bloj, A. Artusi, A. Chalmers, A psychophysical evaluation of inverse tone mapping techniques, Computer Graphics Forum 28 (1) (2009) 13–25.
URL http://wrap.warwick.ac.uk/28433/

[31] L. Meylan, S. Daly, S. Susstrunk, The reproduction of specular highlights on high dynamic range displays, in: Proc. of the 14th Color Imagining Conference, 2006.
[32] M. D. Fairchild, The HDR photographic survey, in: Color and Imaging Conference, Vol. 2007, Society for Imaging Science and Technology, 2007, pp. 233–238.
[33] ITU-R, Proposed preliminary draft new Report - Image dynamic range in television systems, ITU-R WP6C Contribution 146 (April 2013).
[34] S. Daly, T. Kunkel, X. Sun, S. Farrell, P. Crum, 41.1: Distinguished paper: Viewer preferences for shadow, diffuse, specular, and emissive luminance limits of high dynamic range displays, in: SID Symposium Digest of Technical Papers, Vol. 44, Wiley Online Library, 2013, pp. 563–566.
[35] H. Seetzen, L. A. Whitehead, G. Ward, 54.2: A high dynamic range display using low and high resolution modulators, in: SID Symposium Digest of Technical Papers, Vol. 34, Wiley Online Library, 2003, pp. 1450–1453.
[36] A. O. Akyüz, E. Reinhard, Color appearance in high-dynamic-range imaging, Journal of Electronic Imaging 15 (3) (2006) 033001–033001.
[37] S. Kunkel, S. Daly, M. J. Froehlich, Perceptual design for high dynamic range systems, in: F. Dufaux, P. Le Callet, R. Mantiuk, M. Mrak (Eds.), High Dynamic Range Video. From Acquisition, to Display and Applications, Academic Press, 2016, Ch. 15, pp. 391–430.
[38] T. O. Aydın, R. Mantiuk, H. Seidel, Extending quality metrics to full dynamic range images, in: Proc. of SPIE Electronic Imaging: Human Vision and Electronic Imaging XIII, San Jose, USA, 2008, pp. 6806–10.
[39] D. A. Silverstein, J. E. Farrell, Efficient method for paired comparison, Journal of Electronic Imaging 10 (2) (2001) 394–398.
[40] R. Mantiuk, S. Daly, L. Kerofsky, Display adaptive tone mapping, ACM Transactions on Graphics (TOG) 27 (3) (2008) 68.
[41] SIM2, http://www.sim2.com/hdr/ (Jun. 2014).
URL http://www.sim2.com/HDR/
[42] E. Zerman, G. Valenzise, F. De Simone, F. Banterle, F. Dufaux, Effects of display rendering on HDR image quality assessment, in: SPIE Optical Engineering+ Applications, International Society for Optics and Photonics, 2015, pp. 95990R–95990R.
[43] S. Lasserre, F. LeLéannec, E. Francois, Description of HDR sequences proposed by technicolor, ISO/IEC JTC1/SC29/WG11 JCTVC-P0228, IEEE, San Jose, USA.
[44] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, H. Brendel, Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays (2014). arXiv:http://spiedigitallibrary.org.
URL http://spiedigitallibrary.org
[45] ITU-T, Parameter values for the HDTV standards for the studio and for international programme exchange., Recommendation BT.709 (1993).
[46] J. Miles, M. Shevlin, Applying regression and correlation: A guide for students and researchers, Sage, 2001.
[47] R. D. Cook, Detection of influential observation in linear regression, Technometrics 19 (1) (1977) 15–18.
[48] R. D. Cook, S. Weisberg, Residuals and influence in regression.
[49] R. Myers, Classical and modern regression with applications. boston: Pws and kent publishing company (1990).
[50] S. Menard, Applied logistic regression analysissage university, Thousand Oaks, CA.